SN/WA, 31/5-72

PROSJEKTET VARIABELKATALOG OG ET REFERANSESYSTEM TIL ARKIVERT NUMERISK INFORMASJON

av Svein Nordbotten

1. Innledning

Begrepet variabelkatalog har gjennom flere år gått igjen i arbeidsprogrammer og prosjektbeskrivelser. I den senere tid har det bl.a. vært trukket fram i forbindelse med Byråets opplysningsvirksomhet og markedsføring av statistikkprodukter, som hjelpemiddel for å orientere om innholdet av Byråets dataarkiver, utvikling av sosiodemografiske regnskapssystemer, sosialstatistikk, maskinelle framhentingsmetoder, m.m. (jmfr. Eivind Hoffmanns prosjektbeskrivelser, arbeidsprogram og hans notat av 8/2-72 om: Variabelkatalogprosjektet - Foreløpige Synspunkter). Utenfor Byrået begynner flere også å bli opptatt av å lage et system som gjør det mulig å holde oversikt over informasjon som lagres i forskjellige systemer før det hele blir for sent på grunn av det raskt voksende volum. Rasjonaliseringsdirektoratet har derfor tatt initiativ til å få utredet spørsmålet om en registrering av data som finnes lagret omkring i de forskjellige statlige institusjoner. Det utvalg som skal utrede spørsmål om personlighetsvern i tilknytning til offentlige dataarkiver vil naturlig nok også ta opp spørsmålet i en videre sammenheng. Norsk Samfunnsvitenskapelig Datatjeneste, finansiert av NAVF, har som et av sine hovedformål å samle statistiske data og tilby de på en slik måte at de lettest mulig vil kunne utnyttes av samfunnsvitenskapelige forskere. Dette har ført til at NSD har satt i gang og utarbeidd en katalog over statistikk dette organ har arkivert.

Betegnelsen "variabelkatalog" representerer imidlertid ikke noe entydig begrep med alminnelig akseptert innhold og peker i de fleste tilfelle bare på en enkel komponent i et større system. Jeg skal i det følgende prøve å klarlegge hva jeg legger i variabelkatalog og hvordan denne inngår i et referansesystem for de statistiske arkiver.

2. Problemstilling

Når vi arbeider med et samfunnsvitenskapelig problem følger vi ofte en framgangsmåte som skissert i figur 1. I to faser av arbeidet, illustrert ved boks 2 og 4, søker vi å eliminere unødig arbeid ved å gjøre bruk av resultater av arbeid andre tidligere har oppnådd.

När det gjelder fase 2, er framgangsmåten å gå til et fagbibliotek og søke i dets kataloger og indekser etter referanser til eventuelle metoder og løsninger som kan nyttes i det foreliggende problem. Bibliotekenes kataloger er fra gammelt av systematisk oppbygd og vedlikeholdt, og gir et fint grunnlag for søking. Det kan vel føyes til at det for tiden også foregår stor aktivitet med sikte på å forbedre bibliotekenes katalogsystemer og indekseringer.

I fase 4 søker vi etter relevant numerisk informasjon, dvs. statistikk og individualdata. Til tross for at slik informasjon er enklere strukturert og mer standardisert enn tekstlig informasjon, finnes det likevel ikke noe utbygd sidestykke til bibliotekenes katalogsystemer for numerisk informasjon. Når selv ikke Byrået, som må anses som det nasjonale skattkammer for numerisk informasjon om samfunnsforhold, har noe systematisk opplegg for søking og eventuell framhenting av numerisk informasjon, skyldes dette at en tidligere oppfattet all informasjon utenom de trykte publikasjoner på grunn av de betydelige kostnader som framhenting ville medføre, som utilgjengelig i praksis. Den statistiske årsproduksjon som ble gitt ut i 10-20 publikasjoner pr. år, ble videre betraktet som så oversiktlig at det ikke var behov for noe søkesystem for å finne fram til den ønskede informasjon. Etter at moderne databehandlingsutstyr ble tatt i bruk og arkivstatistiske metoder innført, har situasjonen endret seg vesentlig. Tallet på publikasjoner fra Byrået pr. år er nå i størrelsesorden 100, dvs. det utgis 3 000 - 4 000 tabeller pr. år, og et stort antall datasett på lavere bearbeidingsnivå enn publiserte tabeller er lettere tilgjengelig for spesialbearbeiding.

Den måte framleting av numerisk informasjon har foregått på hittil kan kanskje illustreres ved figur 2. En bruker med behov for numerisk informasjon starter vanligvis ved å studere Vegviseren, Publikasjoner fra Statistisk Sentralbyrå, etc. eller ved personlig å konsultere et fagkontor vedkommende mener bør være det riktige å henvende seg til. Dette fører trolig til at det vises til en eller flere publikasjoner. Ved oppslag i tabellregisteret ledes brukeren videre til en eller flere tabeller, og til en eller flere tabellkolonner og -linjer som forhåpentligvis gir relevant informasjon. Ulempene med et slikt system er at det er stor risiko for at det også i andre publikasjoner og tabeller kan finnes relevant informasjon som publikasjons- og tabellnavn ikke avslører. Bortsett fra spredte fotnoter gir systemet heller ikke noen referanse til utrykte informasjonssett, og er derfor utilstrekkelig.

Det ovenfor beskrevne system kan karakteriseres som en systematisk fortegnelse over informasjonssett - publikasjoner og tabeller - som i tabellhoder og forspalter gir referanse til den numeriske informasjon for de begreper (objekttype, objektsamling, kjennemerke, tidsrom) de enkelte informasjonssett belyser. Oversikt over hva som finnes av trykt statistikk om et bestemt emne krever en mer eller mindre fullstendig gjennomgang av alle tabeller. Når det gjelder innholdet i uttrykte datasett som holdes på et eller annet maskinlesbart medium, er situasjonen enda mindre tilfredsstillende. Selv for en byråfunksjonær kan det by på betydelige problemer å få presise opplysninger om innholdet av et gitt informasjonssett. Å få oversikt over hva som finnes om et bestemt emne er i dag ikke mulig.

Det system vi har behov for innebærer at hvert enkelt informasjonssett - publikasjon, tabell, aggregatfiler, og samlinger av individualdata - beskrives på en slik måte at vi kan lage en systematisk liste over de forskjellige begreper som forekommer med henvisninger til de informasjonssett som på en eller annen måte bidrar til å belyse emnet.

3. Et system for å referere arkivert numerisk informasjon

3.1 Systemstruktur

En variabelkatalog er en del av et "information retrieval system". I figur 3 er dette systemet, her betegnet referansesystemet, avgrenset av det prikkede rektangel som igjen kan oppfattes som en del av det arkivstatistiske system representert ved hele figuren.

I det arkivstatistiske system arkiverer statistikkprodusenten numerisk informasjon, individualdata og statistikk, i et data- og statistikkarkiv, F. Denne informasjonen er, eller bør være, nærmere begrepsmessig og operasjonelt definert i et sett med dokumenter - planleggingsnotater, skjemaer, gjennomføringsrapporter, kvalitetsvurderinger m.m. - som er/bør være systematisk satt opp i standardisert form og arkivert tilgjengelig i et dokumentarkiv, P. Referansesystemets sentrale del er en ordnet innholdsfortegnelse, referanseregisteret, R, over hva som finnes arkivert av numerisk informasjon og dokumenterte spesifikasjoner i det arkivstatistiske systemet. Referanseregisteret holdes løpende ajour ved produktbeskrivelser av nye informasjonssett som settes inn i arkivene. Brukere som søker informasjon, formulerer en etterspørselsbeskrivelse og søker etter tilsvarende produktbeskrivelse i referanseregisteret. De funn som gjøres, kommer fram som referansemeldinger med henvisning til hvor i

arkivene det eventuelt finnes informasjon og dokumentasjon av den type som søkes og som gir grunnlag for uttak. Hvor vellykket søkingen vil være, avhenger blant annet av om produsent og bruker uttrykker seg på en noenlunde ensartet måte. For å ta vare på dette krav, må både produsent- og etterspørselsbeskrivelser utformes i et kontrollert, felles beskrivelsesspråk, S.

I det skisserte system er data- og statistikkarkivet den komponent som nå er mest utviklet i Byrået. Dokumentasjonen av innsamling og bearbeiding er varierende, ustandardisert og spredt på forskjellige steder. Oppretting av et systematisert dokumentarkiv bygd på standardisert dokumentasjon av begrepsdrøftinger og -definisjoner, og operasjonelle definisjoner bygd på beskrivelser av innsamling og bearbeiding, bør være en viktig oppgave å løse. Det sentrale skjemaarkivet som foreløpig er det eneste som finnes i systematisert form, kan her danne et naturlig utgangspunkt.

I dette notatet skal vi gå litt nærmere inn på **referansesystemet** og dets komponenter.

3.2 Beskrivelsesspråk

La oss betrakte en gitt produktbeskrevet mengde informasjon symbolisert ved den opptrukne sirkelen A i figur 4. Vi tenker oss videre at den prikkede sirkelen B representerer den mengde som en bruker får henvisning til når han på sin måte lager en etterspørselsbeskrivelse av det han tror er mengden A.

Vi får da tre delmengder:

- mengden (1) er relevant informasjon som bruker ikke får,
- mengden (2) av relevant informasjon som brukeren får,
- mengden (3) av irrelevant informasjon som brukeren får.

Den ideelle situasjon vil være inntruffet om sirklene A og B er sammenfallende. I praksis vil imidlertid produsent og bruker uttrykke seg på forskjellig måte og brukeren vil både få henvisning til irrelevant informasjon og mangle henvisning til relevant informasjon. Forholdet mellom mengdene (2) og (A) betegnes ofte "recall", mens forholdet mellom (2) og (B) kalles "precision". Begge forholdstall vil ligge i intervallet 0 til 1, med 1 som den verdi en søker å oppnå. Byråets nåværende referansesystem som skissert i avsnitt 2, vil ha en relativ lav "recall" verdi.

Målsettingen vil derfor være å lage et beskrivelsesspråk som bidrar til at de to nevnte forholdstall begge blir så nær 1 som mulig.

Forbedring av "recall" kan gjøres ved å karakterisere hvert informasjonssett med flere begreper (f.eks. mer omstendelige publikasjons- og tabellnavn). Forbedring av "precision" oppnås ved å bryte opp hvert informasjonssett i flere og mer homogene sett. Begge framgangsmåter fører med seg økte kostnader. Kostnadsfunksjonen vil vanligvis være slik at med gitte ressurser vil en forbedring av det ene forholdstall føre til en reduksjon i det andre. Uten å ha noen vurderingsfunksjon for de to egenskaper, vil målsettingen derfor i praksis modifiseres til å lage et beskrivelsesspråk som for en gitt "precision" maksimerer "recall", med andre ord, vi godtar en viss prosent av irrelevante referanser, men krever flest mulig henvisninger til relevant informasjon.

Et beskrivelsesspråk vil bestå av to hovedkomponenter, en ordliste og et regelverk for å knytte ordene sammen i beskrivelser. Vi kan tenke ordlisten som sammensatt av fire dellister som representerer "ordklasser":

- liste over objekttyper
- liste over navn
- liste over kjennemerker ved objekter
- liste over tidsspesifikasjoner.

Listen over typer av objekter vil omfatte alle de statistiske enhetstyper som inngår i Byråets undersøkelser, slik som personer, familier, husholdninger, bedrifter, foretak, kommuner, biler, eiendommer, osv. Disse objekttypene angir at beskrivelsen angår et informasjonssett med informasjon om individuelle enheter. Tilsvarende omfatter listen også statistiske aggregatenheter som personaggregater, bedriftsaggregater, osv. og som angir at det pekes på et informasjonssett med statistikk for klasser av personer, bedrifter, osv. For å kunne skille mellom situasjoner når et enkelt objekt beskrives og når en samling objekter beskrives, vil listen omfatte både entalls- og flertallsformer på de forskjellige objekttyper.

Den vanskeligste av dellistene er listen over navn på de enkelte objekter og objektsamlinger.

Navnet beskriver informasjonssettets utstrekning. Vi kan tenke oss navnelisten bygd opp på:

- registre over individualenheter, og
- statistiske klassifikasjoner.

Navnene på de forskjellige registre - inklusive spesialregistre over utvalgsenheter etc. - vil, avhengig av objekttypen:

a. avgrense en samling individualenheter,
b. identifisere en objektklasse.

De enkelte enhetsnavn - identifikasjonsnummer - tillater den enkelte individualenhet beskrevet.

Objekttype: Personer, og navn: Personregister, vil eksempelvis peke på en samling av objekter, mens objekttype: Personaggregat, og navn: Personregister, på den annen side vil angi en statistisk opplysning om bestanden i personregisteret. Objekttype: Person, og navn: xx xx xx xxx xx, vil peke på det enkelte individ.

Navnet på en klasse i en klassifikasjon vil, avhengig av objekttype, enten

a. avgrense en samling individualenheter,
b. avgrense en samling av subklasser,
c. identifisere en objektklasse.

Objekttype: Personer, og navn: Kontorfunksjonærer, objekttype: Personaggregater, og navn: Yrkesklassifikasjon, og objekttype: Personaggregat, og navn: Kontorfunksjonærer, er eksempler på de tre alternative avgrensninger.

Listen over kjennemerker er det vi vel vanligvis tenker på når vi snakker om en variabelkatalog. Denne listen omfatter både de kjennemerker, datakatalogen, vi observerer i tilknytning til de individuelle primærenheter og de kjennemerker, statistikk-katalogen, vi beskriver objektklasser med. Datakatalogen vil også omfatte navnene på våre klassifikasjoner fordi disse også gir uttrykk for kjennemerker ved den enkelte individualenhet.

Variabelkatalogen vil delvis bestå i objektspesifikke kjennemerker. Ekteskapelig status, utdanning o.l. er personspesifikke kjennemerker, mens bruttoproduksjonsverdi og bearbeidingsverdi er bedriftsspesifikke kjennemerker. Andre kjennemerker som alder kan være felles for flere objekttyper. Det er rimelig å tenke seg variabel- eller kjennemerkekatalogen redigert etter kjennemerkenes objekttilknytting.

Listen over tidsspesifikasjoner omfatter alle punkter og intervaller på tidsaksen som nyttes i statistikkproduksjonen. Vi angir punktene på tidsaksen ved åtte siffer, f.eks. 1970.11.01 som tidspunktet for Folketellingen.

Ved å holde en stram kontroll med de betegnelser som innføres i ordlistene slik at det bare er de mest entydige betegnelser som tillates

nyttet, kan dette bidra til å høyne "precision". En måte til å høyne "recall" er å føre inn de vanligste synonymer som likeverdige ord i listene.

Reglene som styrer sammensetningen av ordene i setninger er beskrivelsesspråkets annen komponent. I 1960-årene ble spørsmålet om fritt kontra fast setningsformat inngående drøftet. Fast format er det enkleste og sannsynligvis vil det være en fordel å starte med det. Vi forutsetter derfor at en setning alltid er skrevet slik:

objekttype/navn/kjennemerke/tidsspesifikasjon.

Mellom objekttype, navn og kjennemerke har vi følgende sammenhenger:

Objekttype	Navnetype	Kjennemerke	Angir
Individualenhet			
Entallsform	Individnavn	Individual-kjennemerke	Individual-enhet
Flertallsform	Registernavn Klassenavn	Individual-kjennemerke	Individual-enheter i registeret eller klassen
Aggregatenhet			
Entallsform	Registernavn Klassenavn	Aggregat-kjennemerke	Aggregat-enhet
Flertallsform	Klassenavn	Aggregat-kjennemerke	Aggregat-enheter for subklasser

3.3 Beskrivelser

En beskrivelse av et informasjonssett som arkiveres eller søkes omfatter to eller flere linjer. Hver linje er delt i felter adskilt med tegnet /. Første linje kan vi tenke oss slik:

l.nr. / beskrivelsestype / identifikasjon / tekniske data.

Linjenummeret lar vi alltid være 0 i første linje. Beskrivelsestype er enten Produktbeskrivelse eller Etterspørselsbeskrivelse, identifikasjon kan være navnet på den som har laget beskrivelsen. I Produktbeskrivelsen gir tekniske data referanseadresser til dataarkiv og dokumentarkiv for den lagrede informasjon som beskrives. I Etterspørselsbeskrivelser kan feltet tekniske data spesifisere om det er referanser til dokumentarkiv,

dataarkiv eller begge som ønskes.

De etterfølgende linjer i beskrivelsen er setninger som beskriver informasjonssettet som arkiveres eller lagres knyttet sammen med operasjonssymboler. Formen for hver linje kan vi forestille oss slik:

l.nr. / objekttype / navn / kjennemerke / tidsspesifikasjon / operator.

De operasjoner vi har behov for vil være:

E : logisk "eller",
O : logisk "og",
N : logisk "ikke",
. : stopp.

For å redusere unødig skriving tillater vi gjentakelsestegnet " i felter fra og med tredje linje. Det indikerer at innholdet er det samme som i tilsvarende felt i linjen over.

La oss ved eksempler illustrere hvordan vi kan tenke oss produktbeskrivelser utformet. Et utsnitt fra beskrivelse av personfilen i Folketellingen 1970 kan vi tenke oss slik:

```
0   / Produktbeskrivelse / ......          / ......
.....................................................................
N   / Personer          / Personreg.nov.70 / Fødselsdato/ 1970.11.01/ 0
N+1 /     "            /          "       / Yrke       /      "    / 0
.....................................................................
```

"Personer" er et ord i listen over objekttyper og angir at det er en beskrivelse av en samling personer. I listen over navn, forekommer Personreg. nov. 70 som en avgrensing av samlingen personer. Fødselsdato og Yrke er kjennemerker i variabelkatalogen og 1.11.70 et punkt på tidsaksen. Operatoren O angir at data for både Fødselsdato og Yrke pr. 1.11.70 forekommer koblet i tilknytting til hver person i samlingen. Slik vi her tenker oss Produktbeskrivelsene - en for hvert informasjonssett - vil E og N ikke forekomme i denne beskrivelsestype.

Når Folketellingen er bearbeidd vil det bl.a. foreligge en tabell over antall personer i krysssklassifisering mellom alders- og yrkesklasser. Vi kan forestille oss denne tabellen beskrevet ved:

```
0 / Produktbeskrivelse / ..,..... / .........
1 / Personaggregater / Utdanningsklassifikasjon / Sum / 1970.11.01/ 0
2 /       "         / Yrkesklassifikasjon      / "   /      "     / .
```

Beskrivelsen angir at det er en samling aggregater som beskrives. Målet på hver klasse er Sum personer.

Brukere vil ha behov for å ha et symbol som betegner at innholdet i et bestemt felt er likegyldig. Vi vil bruke tegnet - for dette formål. Vi tenker oss først en bruker som vil undersøke om det foreligger informasjon som gir yrke, kjønn og alder på samlingen av alle personer som var med i Folketellingen 1960 og om dette kan kombineres med deres yrke og inntekt i 1970.

Beskrivelsen vil være:

```
0 / Etterspørselsbeskrivelse / ..... / .....
1 / Personer / Personreg. nov. 1960 / Yrke        / 1960.11.01 / 0
2 /    "     /          "          / Kjønn       /      "     / 0
3 /    "     /          "          / Fødselsdato /      "     / 0
4 /    "     / Personreg. nov. 1970 / Yrke        / 1970.11.01 / 0
5 /    "     /          "          / Inntekt     / 1970.- .-  / .
```

Personer angir at vi ønsker referanse til en eventuell samling av individer som både finnes i Personregister nov. 60 og i Personregister nov. 70. For hvert individ ønsker vi de angitte kjennemerker. At operatoren 0 forekommer her betyr ikke at den søkte informasjon nødvendigvis må foreligge i ett informasjonssett.

En annen bruker er interessert i tallet på døde etter yrke eller etter næring i 1960-årene, og vi kan tenke oss at følgende beskrivelse som illustrerer blandet bruk av operatoren:

```
0 / Etterspørselsbeskrivelse / ..... / .....
1 / Personaggregat / Yrke   / Sum / 196-.-.- / 0
2 /      "        / Døde   / "   /     "    / E
3 /      "        / Næring / "   /     "    / 0
4 /      "        / Døde   / "   /     "    / .
```

Bruk av likegyldighetstegnet betyr i virkeligheten at vi ber om å få referanser svarende til alle de mulige spesifikasjoner som kunne ha stått i feltet.

3.4 Koding

For at beskrivelsene skal kunne behandles effektivt av systemet, må de gjøres mer "behandlingsvennlige" ved en kodeprosess hvor ordene omgjøres til koder.

Kodeprosessen har tre formål:

- å standardisere alternative synonymbetegnelser
- å introdusere relasjoner mellom klasser i et klassifikasjonshierarki
- å effektivisere søkningen

Kodeprosessen foregår ved hjelp av en kodeliste hvor hvert ord i ordlisten har en linje med en kode. Hvert begrep har sin særskilte kode som alle synonyme betegnelser for begrepet har.

Kodens generelle form kan være:

A.B.C....

hvor A er koden for et begrep mens A.B og A.C er koder for underordnede begreper. For objekttyper kan følgende tjene som eksempel:

A : Personer

A.B: Person

C : Personaggregater

C.D: Personaggregat.

Personregisteret er en monotont voksende samling personer, A. Personregisteret nov. 60 er en delsamling med kode A.B, mens Personregisteret nov. 70 er en annen delsamling, A.C, av A.

For klassifikasjoner er dette en velkjent hierarkisk opplysning:

A : Næringen Industri

A.B : Næringsgrenen Næringsmidler

A.B.C: Næringsgruppen Bakerier

osv.

Kodeoppbygningen er viktig for at systemet blant annet skal kunne identifisere subklasser i en setning av typen

objektaggregater / navn / ..... / .....

Dersom navnet har koden A vil de objektaggregater en søker informasjon om ha koder A.x hvor forekommende verdier av x spesifiserer et aggregatobjekt i den søkte samlingen.

Når det gjelder kjennemerker og tidsspesifikasjoner forutsetter vi en tilsvarende oppbygning.

Kodingen kan enten foregå manuelt under utarbeiding av beskrivelsene eller ved maskinelle tabelloppslag på grunnlag av tekstlig utformede beskrivelser.

3.5 Referanseregisteret

Referanseregisteret er systemets sentrale komponent og omfatter alle produktbeskrivelser. Mot dette registeret sammenliknes alle Etterspørselsbeskrivelser.

Vi kan betrakte registeret som en matrise eller tabell som i Figur 5. En kolonne i en slik tabell utgjør en beskrivelse av et informasjonssett, mens hver linje svarer til et begrep. Avhengig av

hvordan en vil søke, kan et slikt register ordnes etter:

- beskrivelse x begreper
- begrep x beskrivelser

Dersom formålet for søkingen er å finne ut hva spesifiserte undersøkelser (representert ved beskrivelser) omfattet, vil den første organisasjonsform være å foretrekke. Det systemet vi hittil har hatt, må i stor utstrekning karakteriseres som et system beslektet med den første typen. Vår problemstilling er imidlertid å finne hvilke informasjonssett (beskrivelser) som inneholder gitte begreper og vi vil derfor foretrekke den andre organisasjonsformen, som kan illustreres ved et system som vist i figur 6.

Ved søking i registeret slås det først opp i tabell B, forspalten, på de begreper som inngår i etterspørselsbeskrivelsen. Hvert oppslag gir en adresse til en linjetabell, L.a., som inneholder en ordnet kjede med referanser til de beskrivelser, S.b, S.c, S.d. ... som omfatter vedkommende begrep. Kjedene kolleres og ferdige referanser ekstraheres i samsvar med etterspørselsbeskrivelsens operatorer.

En produktbeskrivelse legges inn ved at det i tabell B slås opp på hvert begrep som det vises til og at beskrivelsens adressereferanse legges inn i hver linjetabell som oppslagene i tabell B viser til.

Dersom produktbeskrivelsen inneholder et nytt ord kan dette være:

- et synonym til et allerede eksisterende begrep, eller
- en betegnelse på et nytt begrep.

I det første tilfellet må ordet føres inn i kodelisten med koden for begrepet. Når ordet betegner et nytt begrep må dessuten et nytt felt opprettes i B-tabellen med referanse til en ny linjetabell for vedkommende begrep.

3.6 Referansemeldinger

Resultatet av behandlingen av en etterspørselsbeskrivelse er utskrift av en referansemelding. Referansemeldingen skal gi opplysning om:

- hva som finnes av søkt informasjon i dokumentarkiv og data- og statistikkarkiv,
- den form informasjonen finnes på,
- hvor den finnes,
- hvilke betingelser som er knyttet til informasjonen og
- eventuelt hvorfor søkningen ikke har gitt noen referanser.

De fire første kategorier opplysninger er basert på det som finnes i feltet Tekniske opplysninger i første linje i Produktbeskrivelsene. Den siste kategori hentes fra systemet som standardmeldinger på situasjoner som oppstår, som feil i utforming av beskrivelsen, etc.

Når et system av denne type er operativt vil det også være naturlig å tilby individualisert informasjonstjeneste på løpende basis om nye innlegg i arkivene av interesse for brukeren. En slik tjeneste vil bygge på faste etterspørselsbeskrivelser som gir uttrykk for de enkelte brukeres "interesse-profil" og som regelmessig vil bli sammenholdt med referanseregisteret.

4. Oppbygging av et referansesystem i Byrået

4.1 Arbeidsoppgaver og arbeidsprogram

Oppbygging av et referansesystem av den type som ble skissert i avsnitt 3 vil være en omfattende oppgave og sannsynligvis kreve større innsats enn hva de fleste kanskje forestiller seg.

Hoffmann har i sitt notat av 8/2-72 skissert en oppbygging i tre deler. Del I svarer vel stort sett til min ordliste, mens Del II og Del III til beskrivelser for henholdsvis statistikkarkiv og dataarkiv. Hoffmann foreslår at en lar arbeidet med Del II og III gå parallelt etter at Del I er etablert (?).

Jeg vil foreslå følgende framdrift som kan betraktes som en videre utbygging av Hoffmanns forslag.

Fase 1: Begrenset referansesystem

Oppgave 1.1: Ordliste for kjennemerker med referanse til NOS. Arbeidet forutsetter en systematisk gjennomgang av NOS-publikasjonene (begrenset f.eks. til 2-3 siste år) med nedtegning av kjennemerkenavn og tilhørende publikasjons- og tabellidentifikasjoner (jmfr. det som ble gjort ved utarbeiding av norsk-engelsk ordliste), med etterfølgende påføring av "se også" for synonyme eller nesten synonyme betegnelser.

Oppgave 1.2: Ordliste for kjennemerker med referanse til arkiver på maskinmedia. Utgangspunktet vil være filebeskrivelsene med tilleggsinformasjon fra det sentrale skjemaregisteret. Denne ordlisten må samarbeides med den som er nevnt under oppgave 1.1.

Ordnet på en hensiktsmessig måte vil ordlistene med referanser fungere som et begrenset referansesystem. Oppslag i ordlisten vil gi henvisninger til i hvilke publikasjoner/tabeller og i hvilke filebeskrivelser/arkivfiler vi vil finne informasjon om de kjennemerker vi

søker, eventuelt supplert med henvisninger til andre kjennemerkenavn som kan dekke de begreper vi søker å belyse.

En invertering av disse ordlistene vil gi en oversikt over hvilke kjennemerker hver publikasjon og file omfatter (jmfr. kolonnene i figur 5).

Fase 2: Innføring av tidsspesifikasjoner

Oppgave 2.1: Ordliste over tidsspesifikasjoner. Etterhvert som referansesystemet vokser og dekker flere årganger, vil det være ønskelig å kunne skille mellom henvisninger til gammelt og nytt materiale. En ordliste over tidsspesifikasjoner med referanse til publikasjoner og filer lar seg mest hensiktsmessig utarbeide ved å ta utgangspunkt i den inverterte kjennemerkeordliste (dvs. listen over publikasjoner/tabeller og filer) og tidfeste de enkelte kjennemerker som forekommer.

Isolert vil ordlisten med referanser gi oversikt over hvilke informasjonssett som refererer seg til de enkelte tidspunkter eller -perioder. Kombinert med ordlisten for kjennemerker gir den muligheter til å selektere referanser til publikasjoner/tabeller og filer hvor bestemte kjennemerker forekommer for bestemte tider.

Fase 3: Utbygging av et dokumentarkiv

Oppgave 3.1: Dokumentarkivet må bygge på en samling notater i en standardisert form for hver undersøkelse. Prosjektskisser, -beskrivelser, planleggingsnotater, skjemaer, kodelister, revisjonsinstrukser, programhenvisninger, klassifikasjoner, tabellspesifikasjoner, kvalitetsvurderinger, kostnadskomponenter m.m. er komponentene i det som bør utgjøre dokumentasjoner av en undersøkelse og som skal gi brukerne presis informasjon om statistikk og data.

Det må først utarbeides en modell for hvordan en komplett dokumentasjon skal se ut, og deretter må en prøve å få dokumentert de viktigste undersøkelser. Deretter må hver dokumentasjon gis entydig identifikasjon og kjennemerkelisten påføres referanser til disse identifikasjoner.

Fase 4: Etablering av objekttype- og navnelister

Oppgave 4.1: Ordlister for objekttyper og navn. For å redusere omfanget av henvisninger til informasjonssett for objekttyper, objekter og objektsamlinger (øke "precision"), innføres objekttype og navnelister.

Oppretting av navnelistene forutsetter at det foreligger detaljerte beskrivelser av de bestander som er observert eller av de klasser

som inngår i informasjonssettet. Det er derfor rimelig at de opprettes etter fase 3 som vil gi de nødvendige opplysninger.

Fase 5: Automatisering og integrering med arkivsystemet

Oppgave 5.1: Automatisering. Det er rimelig å anta at det gjennom alle faser vil bli gjort bruk av tekniske hjelpemidler i større eller mindre grad, men i den avsluttende fase må referansesystemet automatiseres som et "information retrieval" system for rask reaksjon på forespørsler.

Oppgave 5.2: Integrering. Hittil har vi betraktet systemets output som referanser til den numeriske informasjon vi søker. Det endelige mål vil være å få systemet integrert i det arkivstatistiske system for direkte output av numerisk informasjon.

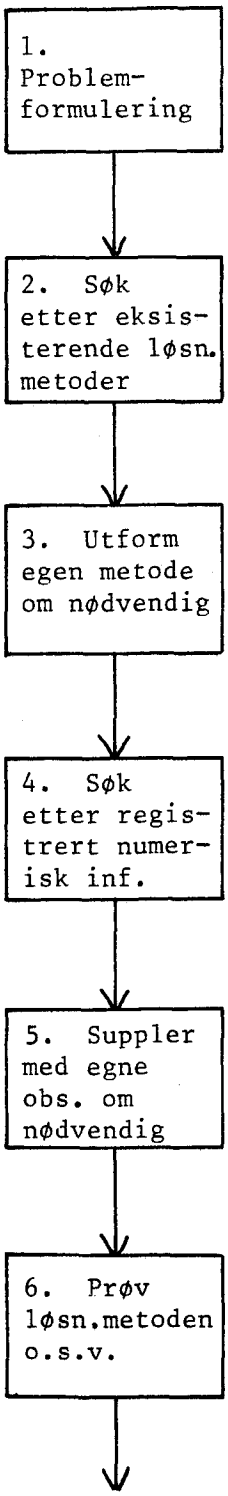
5.2 Arbeidsdeling

Utviklingen av et system som skissert vil forutsette team-work av representanter fra flere grupper. Jeg kan se følgende med sentrale interesser og oppgaver:

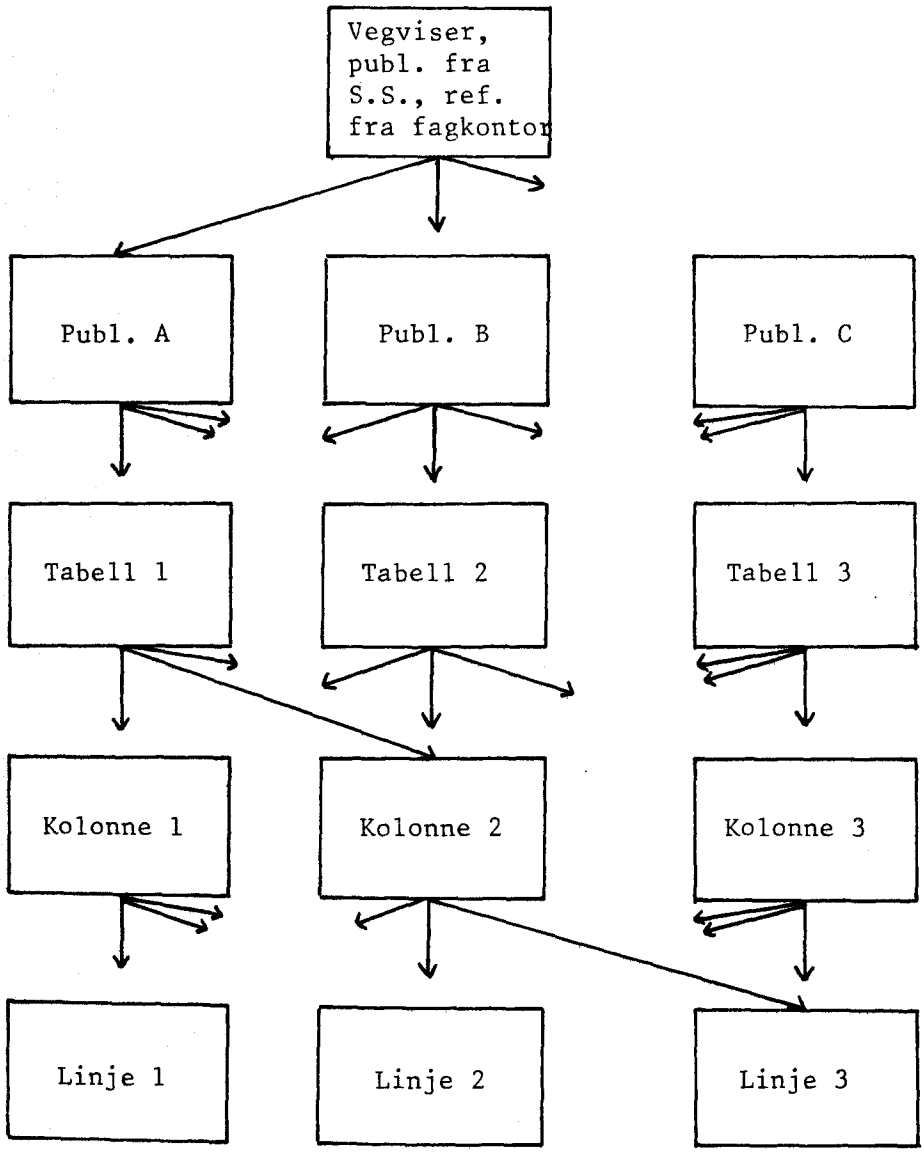
Informasjonskontoret (for Fagavdelingen)
Sosio-demografisk forskningsgruppe
Nasjonalregnskapskontoret (for Forskningsavdelingen)
Systemkontoret (for Produksjonsavdelingen)

Det bør sannsynligvis være to personer, en med faglig-metodisk bakgrunn og en med system-EDB bakgrunn, som samarbeider om prosjektet med en gruppe av representanter fra de nevnte organer som et rådgivende utvalg.

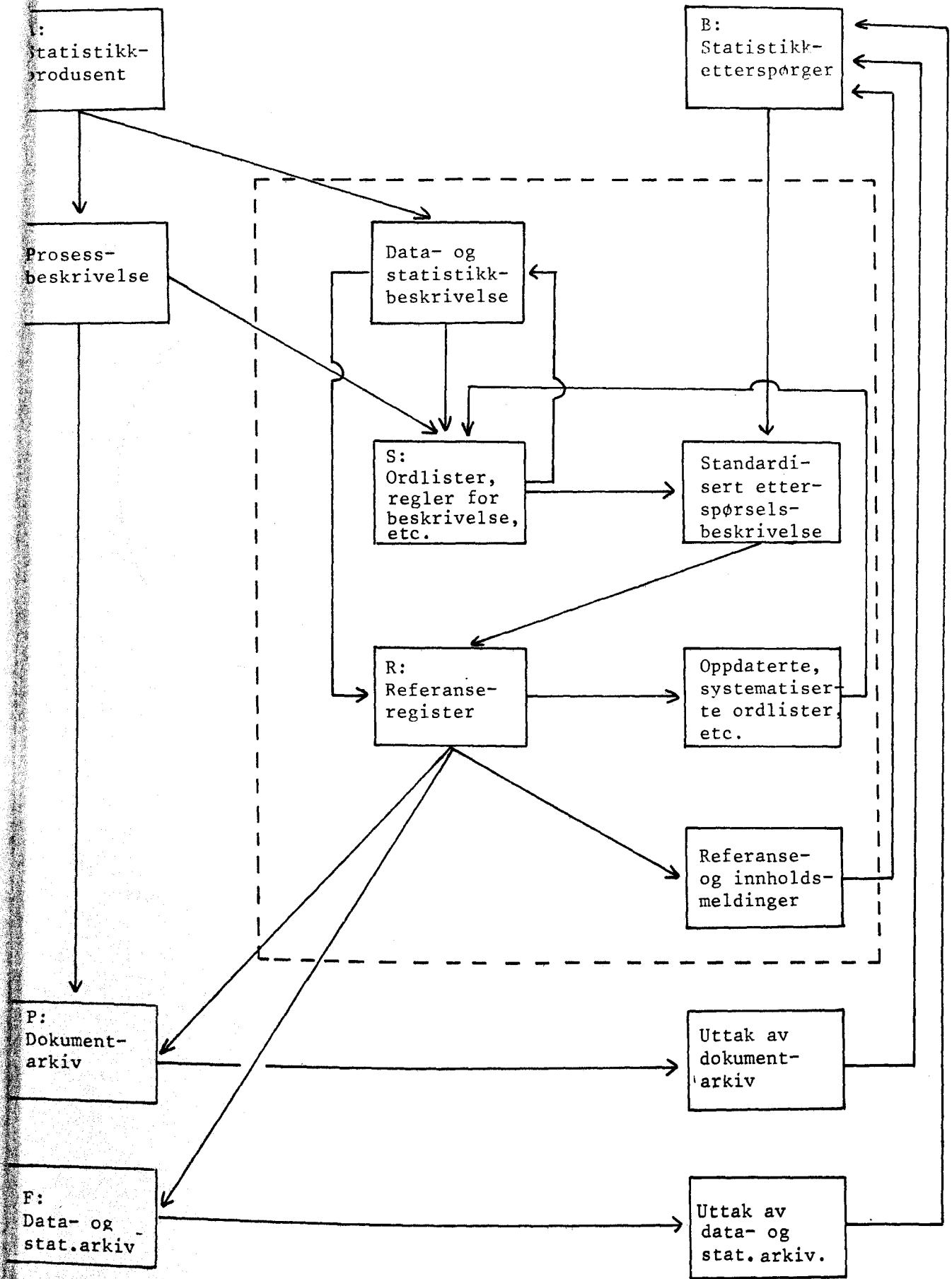
Jeg vil tro det vil være mest hensiktsmessig å legge prosjektet under Produksjonsavdelingen.



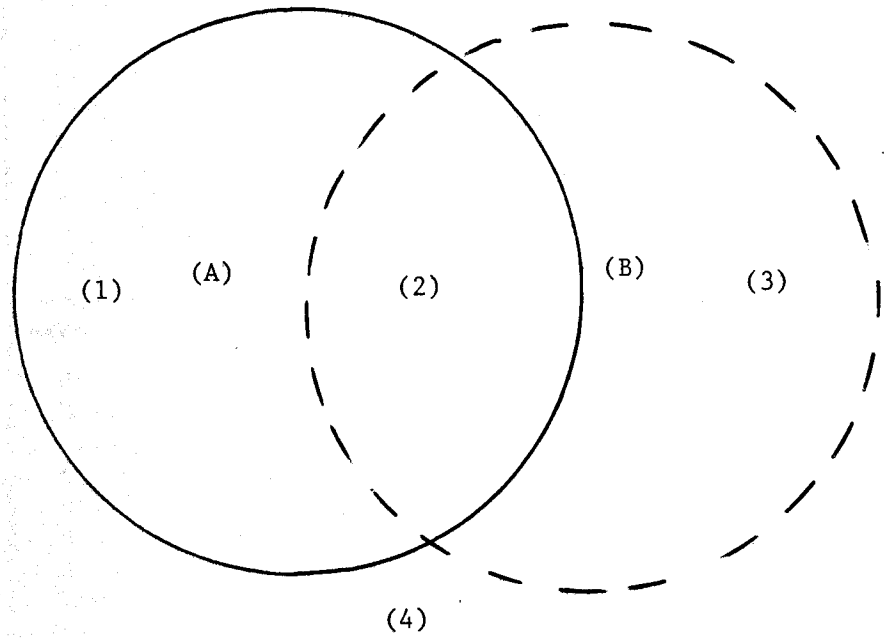
Figur 1: Faser i løsning av problemer



Figur 2: Søking etter statistisk informasjon



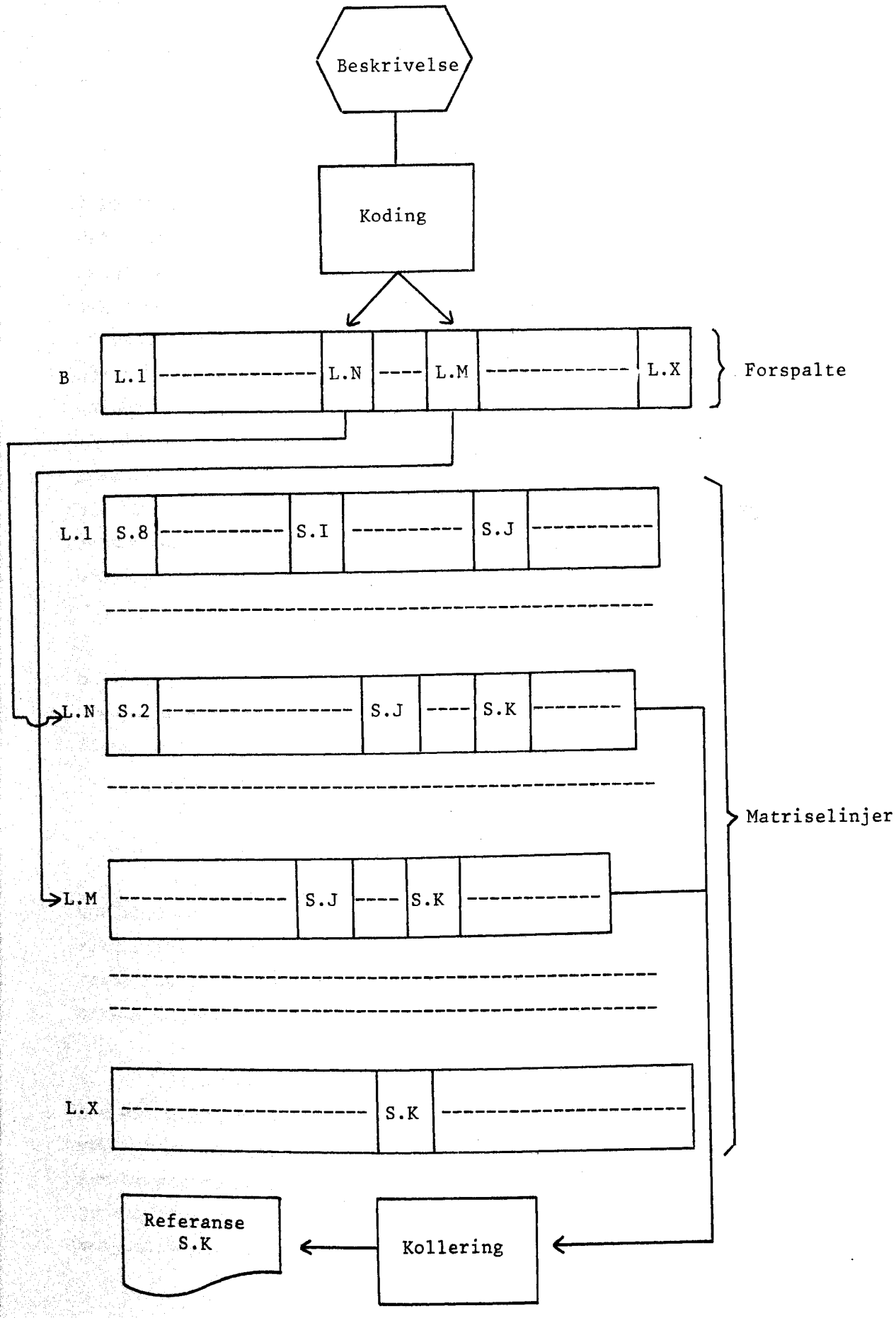
Figur 3: Informasjonssystem for framhenting av data og statistikk



Figur 4: Lagring og framhenting av informasjon

	Beskrivelser						
	Folketelling	Kommuneregnskaper	Arbeidskraftund.	Nasjonalregnskap	Industristatistikk	Data for konsumprisindeks	Pers.reg.
OBJEKTTYPE:							
Personer	x		x				
Familier	x						
Bedrifter					x	x	
Foretak(er)					x		
Kommuner		x					
Transaksjonsaggregater				x			
Bedriftsaggregater					x		
Foretaksaggregater					x		
NAVN:							
Befolkn. i Norge	x						x
Norges kommuner		x					
Utvalg arbeidsaktive			x				
Store bedrifter					x		
Utvalg av detaljomsetningsforr.						x	
Nasj.regn. sektorer				x	x		
Næringsklassifikasjon					x	x	
Kons.vare klassifikasjon						x	
KJENNEMERKE:							
Alder	x		x				
Kjønn	x		x				
Næring	x		x		x		
Utgift		x		x			
Inntekt		x		x			
Aktivitetsstatus	x		x				
Bruttoproduksjonsverdi				x	x		
Gj.pris				x		x	
Pris						x	
TIDSSPESIFIKASJON:							
1.11.1960	x						
1.11.1970	x						
1969		x		x	x		
1970		x		x	x		
4. kv. 1971			x				
1. kv. 1972			x				
Jan. 1972						x	
Feb. 1972						x	

Figur 5: Begrep x beskrivelse matrise



Figur 6: Lagerorganisasjon i et begrep x beskrivelser system